ARI Research Note 2008-08

An Evolutionary Game Theory Model of Revision-Resistant Motivations and Strategic Reasoning

Craig DeLancey
State University of New York—Oswego

Basic Research Unit
Paul A. Gade, Chief

August 2008

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

A Directorate of the Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

MICHELLE SAMS
Director

Research accomplished under contract
for the Department of the Army

State University of New York -Oswego

Reviews by:
Dan Horn, Army Research Institute
Jennifer Solberg, Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Research Note has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPC-ARI-ZXM, 2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926.

FINAL DISPOSITION: This Research Note may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Research Note not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
August 2008	FINAL	October 2004 - to April 2006

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
An Evolutionary Game Theory Model of Revision-Resistant Motivations and Strategic Reasoning	W74V8-05-P-0005
	5b. PROGRAM ELEMENT NUMBER 611102
6. AUTHOR(S) Craig DeLancey (State University of New York–Oswego)	5c. PROJECT NUMBER B74F
	5d. TASK NUMBER 2902
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Office of Research and Sponsored Programs SUNY—Oswego 600 S. College Ave Oswego, NY	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U. S. Army Research Institute for the Behavioral & Social Sciences 2511 Jefferson Davis Highway Arlington, VA 22202-3926	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Research Note 2008-08

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Contractor Officer's Representative and Subject Matter POC: Paul A. Gade

14. ABSTRACT *(Maximum 200 words)*:
Strong reciprocity and other forms of cooperation with non-kin in large groups and in one-time social interactions is difficult to explain with traditional economic or with simple evolutionary accounts. Reciprocity can be costly, while in many instances earning little or no benefit to the individual or its kin. In Ultimatum Games, for example, humans tend in one-shot anonymous interactions towards equal distributions of goods at high individual cost, often encouraged through retributive actions that result in significant personal cost. In this research, an agent-based genetic algorithms model is used to show that in a game similar to the Ultimatum Game, and of which an Ultimatum Game could be interpreted as a subgame, but where the past history of an agent's retributive actions is visible to other agents, strategies exhibiting strong reciprocity can evolve. This model is notable for its conservatism: It presupposes no special features in the structure of the population, relies solely upon potential benefits to kin and offspring, and requires only punishment (and not also reward) as an explanation of the behavior. The model also is consistent with a number of findings on the nature of emotions and related forms of motivation.

15. SUBJECT TERMS
Evolutionary game theory; altruism; strong reciprocity; individual selection; evolutionary theories; behavioral evolution; emotion.

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unlimited	22	Ellen Kinzer Technical Publication Specialist 703-602-8047

Standard Form 298

ARI Research Note 2008-08

An Evolutionary Game Theory Model of Revision-Resistant Motivations and Strategic Reasoning

Craig DeLancey
State University of New York—Oswego

Basic Research Unit
Paul A. Gade, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

August 2008

Army Project Number
611102.B74F

Personnel, Performance and Training

AN EVOLUTIONARY GAME THEORY MODEL OF REVISION-RESISTANT MOTIVATIONS AND STRATEGIC REASONING

EXECUTIVE SUMMARY

Research Requirement:

This report summarizes research carried out pursuant to the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) Contract #W74V8-05-P-0005, under the auspices of its Basic Research Unit.

There is a growing body of research in behavioral economics and experimental economics that shows that human beings often act in ways that fail to be predicted by classical game theoretic and economic models. Many of these behaviors seem to arise because human beings have some kinds of motivations that are not reducible to the economist's traditional notion of ordered preferences. These motivations include some emotions. If we are to be able to better predict the outcomes of strategic interactions ("games"), we will need to better understand the role that these emotions play in reasoning. However, studying the strategic role of emotions is difficult because emotions are hard to generate in scientifically controlled conditions. For this reason, simulations—when designed, revised, and interpreted in terms of human subject research in psychology, experimental economics, and neural science—provide an important method for testing hypotheses and forming general models of the strategic role of emotions. This paper describes a model of the role of anger in reputation building, offering an explanation of the perplexing phenomenon of strong reciprocity. The model is tested through simulations, and then an interpretation based upon empirical findings is offered. The results describe how we can best understand the role of emotions in strategic reasoning: as revision-resistant motivations that offer credible forms of threat.

Procedure:

We used genetic algorithms to simulate the interaction and evolution of strategies in a large population of simple agents. This is a method that allows for the testing of hypotheses in evolutionary game theory that would otherwise be intractable because of their complexity. The simulation is designed in light of the most current human subject research on a widely studied game: the Ultimatum Game. This allows us to test the hypothesis that human Ultimatum Game performance is best explained as a special instance of performance in another more general class of game, the Reputation Game.

Findings:

The simulations confirm that an evolutionary stable strategy for the Reputation Game approximates the performance of humans in Ultimatum Games. We introduce a new game theory concept—the quasi-subgame—as a way to explain how an optimal strategy for the Reputation

Game applies to the Ultimatum Game. This finding also is consistent with the interpretation that Reputation Game performance is best explained in individual agents like humans as a revision-resistant inherited motivation, such as anger. This approach assumes that our ancestors were frequently in strategic interactions where their reputation as potential retributors could be known and play a role, and rarely in conditions like those in an Ultimatum Game. This is a strong consequence and one that makes the simulation amenable to falsification.

We also offer an interpretation at the level of the individual organism of how these strategies are instantiated in specific forms of motivation. This is the hypothesis: that some emotions are revision-resistant preferences that motivate a kind of behavior.

Utilization and Dissemination of Findings:

This research has the potential to improve predictions of strategic interactions where anger or a reputation for retribution can play a role. Such interactions are likely common in the conditions Soldiers face and that those evaluating and developing strategy must consider.

AN EVOLUTIONARY GAME THEORY MODEL OF REVISION-RESISTANT MOTIVATIONS AND STRATEGIC REASONING

CONTENTS

LIST OF FIGURES

Background

This report defends the hypothesis that human beings, and likely other animals, have revision-resistant motivations that play important strategic roles in decision making. Emotions like anger, fear, and disgust are primary examples of such motivations. A model in evolutionary game theory, tested through simulations, provides an illustration of how some revision-resistant motivations could evolve, and also both explains such behavior as altruistic punishment and provides predictions suitable for future empirical research. The interpretation of the simulation uses a novel form of game-theoretic explanation: approximate case games. The model and its interpretation are consistent with our best scientific understanding of emotions. A conceptual discussion using game theoretic concepts clarifies the potential individual benefit of revision-resistant motivations and why this concept should not be replaced with the concept of social preferences or other standard economic notions of motivation.

There is a growing body of research in behavioral economics and experimental economics that shows that human beings often act in ways that fail to be predicted by classical game theoretic and economic models. Examples include human subject research on Ultimatum Games (see Camerer, 2003, pp. 50-55), Public Goods Games (Andreoni, 1995; Gintis, Bowles, Boyd, & Fehr, 2003), and a range of forms of limited rationality in individual reasoning (see Camerer, 1995). Many of these behaviors in cases of strategic interaction appear to arise because human beings have some kinds of motivations that are not reducible to the economist's traditional notion of ordered preferences. Such motivations include some emotions, which appear to be resistant to revision in the ways that preferences can be revised. Better understanding and predicting strategic interactions where emotions can play a role will require an improved understanding of how these emotions shape strategic interactions ("games"). This paper describes a model of the role of anger in reputation building, offering an explanation of the perplexing phenomenon of strong reciprocity. The interpretation of this model, based upon empirical findings, describes how we can best understand the role of emotions in strategic reasoning: as revision-resistant motivations that offer credible forms of threat.

Simulation of Retribution in Approximate Games

Many forms of human behavior appear to demonstrate altruistic cooperation or altruistic punishment. These are difficult to explain in simplistic evolutionary models because (by definition) such behaviors appear to provide no benefit to the individual or its kin. Human performance in Ultimatum Games is one example of such behaviors, where in one-shot interactions with non-kin strangers many individuals are willing both to share and to punish altruistically (Guth, Schmittberger, & Schwarze, 1982; Camerer & Thaler, 1995). Because these interactions are anonymous and one shot, they appear not to be well explained by kinship selection (Hamilton, 1964), nor by reputation building and signaling (Nowak & Sigmund, 2005). Ultimatum Game performance and related kinds of seemingly altruistic behavior can be explained through one form of reputation building, given that the kinds of interactions seen in the Ultimatum Game can be interpreted as an extension of strategies adopted for a more general class of interactions that include the opportunity for reputation building and its benefits to self and kin. Here the reputations in question are not for cooperation (as in image scoring—Nowak & Sigmund, 1998) but rather for a willingness to punish at personal cost. I make this argument by

describing a simulation that tests two hypotheses about reciprocal behavior in a game that shares some features with the Ultimatum Game. First, I propose that the ability to develop reputations as retributors leads to the evolution of a high frequency of adoption of a retributive strategy, resulting in more equitable offers in games with those who have developed such reputations. Second, I propose that the presence of retributors will lead to more equitable offers in all game rounds played among a closed population, including in game rounds played with those with no reputation for retribution. It is only the second hypothesis that could plausibly explain the behavior of humans in Ultimatum Games, since more equitable distributions are seen even in situations where proposers have no reputation for retribution.

The details of the Ultimatum Game are now widely discussed but a brief review may be helpful. In an Ultimatum Game, there are two players, the responder and the proposer. These players do not know each other or interact except through their moves in the game. The proposer is given a sum of utility that we can for simplicity assume is divisible into discrete units (say, thirty-two $1 bills). The proposer has to decide how to split this sum with the responder. Both the proposer and responder are aware of the size of the utility, and fully aware of the rules of the game, including that it is anonymous and is played in single rounds (that is, they know any future game is unlikely to be played with this individual). The proposer can offer any amount; for simplicity we can also assume that the proposer will offer only non-zero amounts (such as $1, $2, up to $32).[1] The responder can then accept or reject the offer. If the responder accepts, the money is divided as proposed. If the responder rejects the offer, neither player gets anything. Whether the offer is accepted or rejected, the game is then over. (In what follows, we will refer to this particular instance of an Ultimatum Game as "the Ultimatum Game.")

Both standard game theory and standard economic theory, in which one assumes that agents seek to maximize subjective personal utility, predict that the proposer in the Ultimatum Game will offer the minimal amount, and the responder will accept this amount (Rubinstein, 1982). Depending upon how one sets up the game, there are typically many Nash equilibria (for example, as we set up the game in our simulation below, there can be 528 Nash equilibria).[2] But in all standard Ultimatum Games where we allow only discrete proposals greater than zero, there is only one subgame perfect strategy: For proposers it is to offer the minimal amount, and for responders it is to accept any offer, including the minimal offer (see Selton, 1965; Gintis, 2000). This makes for an intuitively plausible prediction; the proposer can maximize its utility by

---

[1] The assumption that the utility is divided into discrete chunks allows there to be a subgame perfect equilibrium in the game, and is realistic since money (the usual form of utility used in human subject experiments) does not come in continuous magnitudes. The assumption that a zero offer is not allowed ensures a single subgame perfect equilibrium by ruling out the case of a zero offer, which in standard game theory would be seen as an indifferent case for the responder and so count as a kind of trivial or vacuous subgame perfect equilibrium. Both assumptions simplify the account but should not alter the interpretation of the empirical findings.

[2] We use a simplified version of the Ultimatum Game in which responders come to the game with some minimal acceptable offer expectation, and will reject offers below that. Thus, when a proposer offers some utility N, any minimal acceptance value for a responder of N or less is a Nash equilibrium strategy.

making the smallest possible offer; while the responder recognizes that some offer, however small, is better than nothing, and so can maximize utility by accepting any offer.

If the utility in question benefits fitness, then evolutionary theory also would suggest that one-time interactions with non-kin should result in games exhibiting the subgame perfect equilibrium. The proposers more likely to have offspring are those that maximize fitness, and they can do this by offering the minimal amount; the responders are more likely to have offspring if they maximize fitness and can do this by accepting any offer. A more equitable distribution (that is, something like an even sharing of the utility) in such interactions cannot in any direct or obvious way benefit the proposer or the proposer's kin. Since by supposition in the standard Ultimatum Game these two agents are not kin and are unlikely to ever meet again, it would seem that the proposer should offer the minimal amount. And even if the responder's ability to reject is interpreted as punishment or retribution, it appears that the responder receives no benefit from punishing an agent that he is unlikely to meet again. Given that the punishment is costly, it would appear better to simply accept any offer made.

Actual human behavior in the Ultimatum Game, however, is strikingly different than this prediction (for a summary of findings, see Camerer, 2003, pp. 50-55). In most cultures, proposers tend towards equitable (that is, equal) divisions, and responders are surprisingly likely to reject lower offers. Proposers tend to offer a mean of nearly 40% of the stakes, and responders frequently reject low offers, such as those around 20% of the stakes or less. This finding varies significantly across cultures, but nowhere do we see the subgame perfect strategy being pursued (Henrich et al., 2004; Henrich et al., 2005).

Ultimatum Game play thus appears to exhibit what Bowles and Gintis (2002) have termed *strong reciprocity*: cooperative or altruistic behavior at significant personal cost that arises even when it is implausible to expect that those costs will be (directly) repaid. These findings provide a significant challenge to the more simple forms of standard economic, game theoretic, and evolutionary explanations.

In what follows, an evolutionary game theory model of Ultimatum Game performance is offered in which Ultimatum Games are treated by players as a particular instance of a more general game, which we can call the Reputation Game. In this more general game, agents have the ability to develop a reputation for willingness to punish, and this results in the evolution of a proposer strategy of making more equitable offers and a responder strategy of rejecting low offers. Ultimatum Game performance is explained as an application of this general strategy. The preferred interpretation of the model depends upon benefits to self and to kin. However, the performance of any one individual may be independent of any particular benefit they or their kin may receive, so that strong reciprocity is observed even in anonymous one shot games.

Reputation Model in Simulation

The guiding assumption of this simulation is that Ultimatum Game performance is a product of a strategy that evolved or was adopted in an environment in which agents are involved in interactions very similar to those in the Ultimatum Game except that most of these interactions are not anonymous. In such an environment, an agent can develop a reputation for being a

punisher or retributor. This reputation, in our model, is then potentially beneficial to the individual and the individual's offspring. Some agents can be known by other agents as being willing to reject offers even when this results in a substantial personal loss. One can expect knowledge about these reputations, among proposers, to raise the mean offer in a population towards more equitable distributions for those retributors. However, one can expect that the population as a whole then comes to benefit, as the benefits of being willing to reject low offers begins to increase the benefits for being willing to reject offers (given merely the potential that one could come to be known as a retributor), so that even those who do not develop a reputation for retribution begin to require and receive higher offers.

To test this, one can introduce a Reputation Game. The Reputation Game is like the Ultimatum Game, except that the responder has a reputation that the proposer can consider in making an offer. In extensive form, the Game has three basic moves: The responder communicates its history of rejections, the proposer makes an offer, and then the responder accepts or rejects that offer. I used a simulation of a population of simple agents—each with five parameters—that play this game. Four of these parameters are coded as five-bit binary strings, interpreted as representing values of 1-32 with standard binary interpretation (and the addition of 1). Each of these four parameters freely evolved (that is, the first generation had random values for each bit value for each individual, and then each generation is allowed to compete to have "offspring," using standard genetic algorithm approaches of sexual reproduction with mutation). These parameters are primary proposal, rejection-count threshold, secondary proposal (these three parameters apply when the agent is a proposer), and minimal acceptable proposal (the sole relevant evolving parameter when this agent is a responder). The fifth parameter is the rejection count, which is a simple integer: This value begins as zero, and during a mock game play phase of thirty-two games as a responder, each time an agent rejects an offer, this value is incremented. This can be thought of as the agent's reputation, announced at the beginning of a round, when the agent is playing the role of responder.

During the actual game play phase, games are played by each agent as a proposer, against some randomly selected other agents in the population. In each game, the proposer can see the rejection count of the responder. If the responder's rejection count is greater than the proposer's rejection-count threshold, the proposer will use its secondary proposal. Otherwise, it uses its primary proposal. This proposal is then communicated to the responding agent. When playing as a responder, the agent will reject any offer not greater than or equal to that agent's minimal acceptable proposal value.

Agents play 32 games as the proposer each generation, and responders are selected randomly from the population, so that on average each agent will also play 32 games as a responder and each is very unlikely to be chosen twice by the same agent. Games are thus one-shot, though not anonymous. Reward in the game is incremental fitness, which is summed across the games played at the end of all game play for each generation. Agents are then selected using roulette selection (Mitchell, 1998), a procedure that makes the likelihood of reproducing proportional to fitness (thus allowing for some small chance that seemingly unfit strategies will reproduce). Reproduction is sexual, using a single random cross-over point with another roulette-selected agent. Each offspring had a 1% change of having either (half the mutations) a random 1 or 2 bit mutation (the flipping of a bit) at a random point in the bit string that constitutes the four

evolving parameters; or (the other half of the mutations) a random ±1 change in the interpreted value in one of those parameters.[3] Each agent survives only one generation; this is the reason for the round of "mock" games played before actual games, in which the incremental fitness reward is not actually distributed but rejections of offers are counted to increment each agent's rejection count parameter (the rejection count parameter is not incremented during the actual game play phase). Simulations started always with a random population of 5,000 individuals, and evolved over 5,000 generations.[4]

Simulation Results

The simulation confirms that reputation for punishment leads to a more equitable distribution of utility in games played with those who have a reputation for retribution. This is evident in the fact that the second proposal value, which is the proposal value for agents who have shown a history of retribution, settles on a population mean value of 9.2, with a population standard deviation staying close to 0.5. Figure 1 shows the mean population primary and secondary proposal values for 5,000 generations, normalized to range from 0 to 1. The typical population, after 300 generations, has a mean primary proposal of 9.0 out of 32 but a wide standard deviation of 6.7. The population mean rejection threshold settles at 3.8 out of 32, and the population mean minimal accept value settles near 6.8 out of 32.

Most noteworthy is that the results confirm the second hypothesis and show that primary proposals also evolve towards more equitable distributions in line with human performance—although with a wide variation. A simplistic expectation for the simulation might be that agents will evolve the lowest possible primary proposals, but higher secondary proposals. This way, they could make equitable proposals to retributors but low proposals to those without a reputation for retribution. Instead, the population means for both the primary and secondary proposals quickly evolve towards a more equitable value. Our explanation is that as agents begin to evolve high secondary proposals, it becomes in the interest of agents to develop a reputation over time for rejecting low proposals. This creates the opportunity for there to be offspring with higher minimal accept values. This also in turn creates a general environment where offspring are facing fewer low proposals, and so many agents with higher minimal accept values will not reject many or any proposals and will not then develop a reputation for retribution (their rejection-count will be low or zero). The result is that reputation building helps not only agents that reject frequently but also benefits the entire population. The reputation for rejecting low proposals, in other words, starts to cast a halo over the whole population; even those that have made very few or no rejections come to have higher accept values and other agents make higher primary proposals because their ancestors found it beneficial to avoid rejections. Agents have an "interest" in testing the rejection threshold of other agents, and this explains the wide standard deviation.

---

[3] This latter form of mutation is included to avoid local maxima that can result as an artifact of binary number coding.

[4] Tests of populations beyond 5,000 generations or with populations larger than 5,000 agents showed no significant variation from the values already observed.

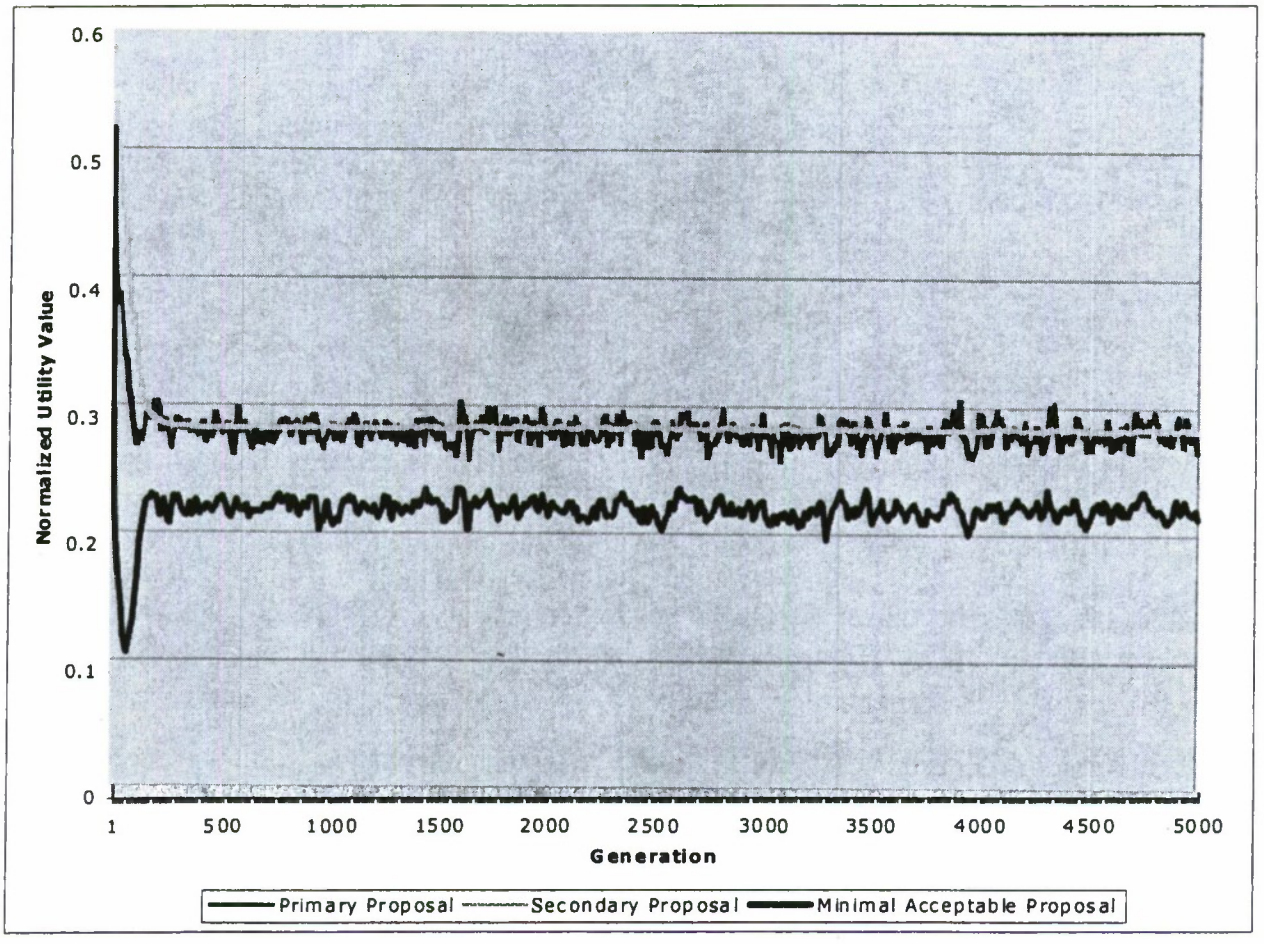

Figure 1. *Population mean values for Reputation Game.*

Our proposal is that an Ultimatum Game could be seen as a kind of approximate instance of this Reputation Game. The Reputation Game is one shot but not anonymous. However, an anonymous interaction is an interaction where the proposer knows nothing of the responder's rejection history. If we interpret this as the anonymous responder having, from the perspective of the proposer, a rejection history of zero (or, alternatively, *any* rejection history), then the proposer will be offering a mean value around 9 out of 32. This is 28% of the utility. Camerer's summary of Ultimatum Game findings shows a mean proposal typically ranging from 30% to 40% (Camerer, 2003). Furthermore, the population mean minimal acceptable proposal value tends towards 21% of the utility. Camerer's summary found that offers below 20% were rejected about half the time, whereas offers closer to 40% were rarely rejected. Thus, if we interpret Ultimatum Game performance as an application of the dominant Reputation Game strategy to the Ultimatum Game, these findings are close to human performance.

Approximate Instance Games

This model has on its side both plausibility and parsimony. It is plausible that human beings and their ancestors often interacted with the same individuals in non-anonymous situations, and that reputation building can play a role in the social benefits of retribution, so that

something like the Reputation Game is more common and more important to humans than the Ultimatum Game. The strategy that agents use in Ultimatum Games would thus be a strategy that evolved or developed in response to something like Reputation Games, and then this strategy is applied in Ultimatum Games because (in a sense discussed below) Ultimatum Games can be interpreted as an instance of Reputation Games; that is, there is a strategy in the Reputation Game that can be interpreted as a strategy for an Ultimatum Game—in our case, interpreting the lack of retribution history as equivalent to a retribution history of zero (or any arbitrary value). Thus, the model can offer an explanation of human Ultimatum Game performance with no additional hypotheses.

If we describe the Reputation Game in normal form, then it will have a finite set of players: a set A_{Ri} of action for each player i, and a preference relation on those actions for that player. Similarly, an Ultimatum Game will have a finite set of players, a set A_{Ui} of actions for player i, and a preference relation on those actions for that player. But for the proposers, the actions required are similar; make an offer. For the responder, there is an additional action in the Reputation Game—announce one's history, but otherwise the same action of accept or reject the offer. Thus, for both proposers and players, it is the case that (as the games have been set up here, with the same absolute values, and the same relevant constraints on those values):

$$\forall x\ (x \in A_{Ui} \rightarrow (\exists y\ y \in A_{Ri} \wedge y \subset x))$$

That is, every action of the Ultimatum Game is the subset of some action of the Reputation Game. Furthermore, every possible strategy for the Ultimatum Game can be had in the strategies for the Reputation Game (no strategy of the Ultimatum Game is inaccessible to the set of strategies for the Reputation Game, if the Ultimatum Game is interpreted as an approximate instance of the Reputation Game). Similarly, if we describe the games in extensive form, then the Ultimatum Game includes a set of sequences of moves that is a possible history set, S_U, as does the Reputation Game, S_R. Each sequence of S_U will be a subsequence of some sequence in S_R (that is, a history in S_U is like a history in S_R with some steps removed). Let us say that a game, B, can be seen as an "approximate instance" of another game, A, when either of these conditions occurs. Intuitively, this means that all the moves of game B can be treated as part of game A, so that a strategy for game A could be applicable as a strategy for game B by ignoring the actions or sequence steps that are not relevant to game B (this says nothing about whether any strategy should be preferred in game B if it is preferred for game A).

Our explanation of human Ultimatum Game performance thus introduces a novel principle: the interpretation of a game as an instance of a more common game for which the agent has an existing strategy. Nonetheless, an Ultimatum Game is distinct from the Reputation Game. Explanations built upon this notion must make a plausible case that the benefit of an alternative strategy or of adopting a mix of several strategies is less than that of an "extended" strategy.

We can conceive of this more clearly if we imagine that the agents are acting in a population where the other agents they encounter may play with them in one of two games—A or B, where B is an approximate instance of A. This includes the assumption that the set of strategies S available to the agent includes optimal strategies for both A and B. For example, a

strategy equivalent to the subgame perfect equilibrium strategy of the Ultimatum Game is available to the agents with the resources of the strategies for the Reputation Game (for example, both primary and second proposals could be 1). We suppose that some portion N of the population plays A, and the rest play B. Let S_1 be an optimal strategy for playing A, and S_2 an optimal strategy for playing B (as described in the strategies for game A). Let $U_A(S_1)$ be the mean utility of playing strategy S_1 in game A. Then, the expected utility E_P of playing strategy S_1 in the population is expressed by:

$$E_P(S_1) = (U_A(S_1)*N) - (U_B(S_1)*(1-N))$$

This assumes that there is no additional cost to developing one strategy over another, or in choosing between strategies. If S_1 and S_2 were the only available strategies, then S_1 should dominate S_2 when $E_P(S_1) > E_P(S_2)$. If a range of strategies is possible and there is a linear relation between their relative benefits, we should expect the dominant strategy to approach S_1 as $E_P(S_1)$ exceeds $E_P(S_2)$.

We might suppose, however, that the organism has the resources to develop and use two strategies in the games that it plays in the single population. If there is some incremental cost that is spent in having the resources for an additional strategy, call this mean cost C_1. This might be the cost of additional developmental resources devoted to this additional strategy, or of the maintenance and use of that strategy. Furthermore, there can be some cost in distinguishing between the cases. Previously, we assumed that the agent played a single strategy in every interaction. The alternative is that it must distinguish the appropriate kinds of game being played by the other player in order to act with the optimal strategy. Call this mean cost C_2. It might be a cost in time and energy required to find and interpret evidence. Then—again assuming for simplicity that there are only two available strategies, S_1 and S_2—S_1 would dominate the alternative of having two strategies, S_1 and S_2, when

$$E_P(S_1) > (U_A(S_1)*N) + (U_B(S_2)*(1-N)) - C_1 - C_2$$

We have proposed that humans and their ancestors have consistently been far more often in strategic interactions more like the Reputation Game than the Ultimatum Game. If we grant that an Ultimatum Game can be an approximate instance of a Reputation Game, then we also are assuming that the Reputation Game is so much more common, or the costs of maintaining and switching between other strategies is so great, that the extension of a strategy suitable for the Reputation Game is typically found in the Ultimatum Game.

We illustrate this relationship in our simulations by assuming C_1 and C_2 are negligible, and exploring the proportionate range of Reputation Games (here representing game A) with Ultimatum Games (here representing game B). In Figure 2, the x-axis shows the relative proportion of Ultimatum Games to Reputation Games in the set of games played by each agent, in 5% increments. The y-axis shows population mean primary proposals and minimal accept values after 5,000 generations; again, the values ranged from 1 to 32 but are here normalized as a range from 0 to 1. The Reputation Game parameters were used, but the relative proportion of Ultimatum Games (stochastically selected) were treated as games in which only the primary proposal parameter and the minimal accept value parameter were utilized and an Ultimatum

Game was then played (in which the primary proposal was rejected if below the minimal accept value). As predicted, the primary proposals fall towards the subgame perfect equilibrium of minimal proposal and accept values, as the proportion increases. The relationship is relatively linear and smooth. (The primary proposal value for 100% Ultimatum Games is slightly above the subgame perfect equilibrium because of the noise introduced through the mutation rate and the relative benefit agents receive from keeping the primary proposal above the minimal acceptable proposal value.)

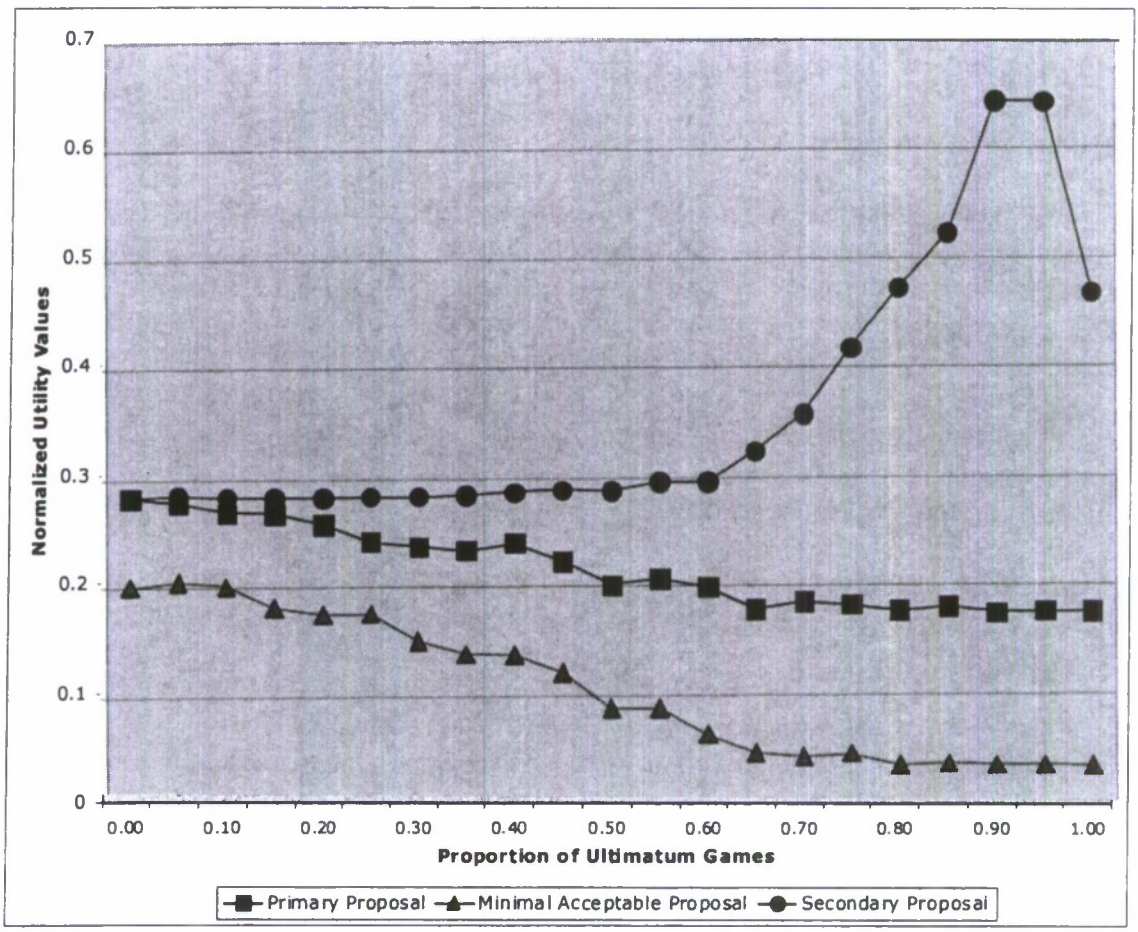

Figure 2. *Reputation Game and Ultimatum Game mix.*

The secondary proposal has a narrow standard deviation (less than 2) until the proportion of Ultimatum Games exceeds 65%, when the mean starts to climb sharply. Then the standard deviation approaches that of a random value, and in fact the final value for the secondary proposal settles near 50% of utility because there is a random distribution of secondary proposal values. This suggests that as retributors become rarer, the best strategy is to pay them off with a high offer, minimizing the risk of a rejected offer. Also, as the proportion of Reputation Games shrinks, those agents willing to reject offers tend to be mutants with high minimal acceptable proposal values.

This illustrates the underlying assumption of the model. To the degree that Reputation Games are more common than Ultimatum Games in the population, the strategies will reflect this, and more equitable proposals and more demanding minimal acceptable proposal values will occur in the population, approaching the optimal solution of proposals (both primary and secondary) settling near 28% or more of the utility in the Reputation Game, with the minimal acceptable proposal value approaching 21%. If behavior of humans in the Ultimatum Game is best explained by strategies acquired in playing Reputation Games, the prevalence of interactions like Reputation Games must have been far more common than interactions like Ultimatum Games, or the cost of switching to an alternative strategy, or maintaining that alternative strategy, must exceed the benefits to be gained.

Emotions as Strategic Motivations

The kind of evolutionary game theory explanation simulated here is neutral regarding whether it describes a strategy that is inherited or culturally transmitted, but depends upon the presumption that agents acquire and retain a strategy in a way resistant to immediate learning. The ancestors of human beings lived in socially rich environments where most strategic interactions were with known individuals with whom one was likely to interact again in the future. It is for this reason that we assume that the Reputation Game is much more representative of a typical human interaction than is an Ultimatum Game. The model is neutral regarding whether the evolution of the strategies is realized through inherited structures, or cultural transmission of behaviors, or any other form of transmission. Rather, all that is supposed by the model is that strategies are repeated and variations of them attempted, and the strategies that are reproduced tend to be those that beat out other strategies. Consistent with this form of explanation, once a strategy is adopted by an agent, it may be the sole strategy that agent uses throughout all game play in the Reputation Game, and similar games or approximate instances. That is, there is no reason to suppose that the evolution of strategies is occurring within any particular agent; the more simple application of the model is that each agent is using one strategy. This allows the model to explain behavior in approximate instances; immediate learning is not a factor in the explanation.

This explanation is consistent with a growing body of evidence concerning human emotions. One way to interpret the psychology of Ultimatum Game performance is that agents become angry, and anger motivates retribution. One recent study shows that brain regions differentially active during anger are also differentially active in individuals rejecting proposals in an Ultimatum Game when playing as responders (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). This is precisely what we would expect if anger were a motivation to retribute in the Reputation Game, and the display of anger or the reputation for anger was the way to announce willingness to retribute; and if it was then the case that anger acted as the motivation in the approximate instance of the Ultimatum Game. A similar form of explanation utilizing anger seems appropriate for other similar kinds of altruistic retribution, such as in Public Goods games (see Andreoni, 1995; Gintis et al. 2003).

An important feature of anger and other motivations is that many of our emotions appear revision-resistant. That is, unlike the kinds of preferences that economists typically study, they cannot be easily altered. For this reason, a number of economists have argued that emotions can

act as special commitment mechanisms, enforcing behavior that would otherwise be less likely to occur (Frank, 1988; Hirshleifer, 1984). Consistent with this, neural scientists have uncovered a growing body of evidence that some emotions and other forms of preferences are significantly independent of our abilities to consciously process information (Zajonc, 1968, Kunst-Wilson & Zajonc, 1980; Öhman, 1988; Öhman, Dimberg, & Esteves, 1989; Öhman & Soares, 1993, 1994). An inherited pancultural disposition could behave in this way, and remain, because its resistant to revision was not harmful or was even beneficial. Finally, many species of organisms have invested significant resources into the ability to display threats, suggesting that the role of making credible threats is an important survival skill. Organisms find it worthwhile to expend energy to create displays meant to convey their ferocity and willingness to fight. Lions growl, dogs bare their teeth, birds inflate themselves to exaggerate their size, and so on. These displays may be interpreted as attempts to create and convey a reputation of being willing to retribute, avoiding some conflicts but also, in some cases perhaps, earning more equitable treatment from other conspecifics. These findings are consistent with the idea that anger acts as an inherited disposition to retribute, that we benefit from communicating this disposition, and that we may act on this disposition even in cases where the benefits of reputation are lost because the disposition is not easily revised or controlled.

Conceptual Discussion of Revision-Resistance

Economists and game theorists typically aim to explain human action by supposing the agent is rational, has beliefs (information), and has ordered preferences. Whether an agent is seeking to maximize his preferences, and whether those preferences need to be selfish, are matters of contention (see Sen, 2002). However, most such explanations share a conviction that the agent acts in part to satisfy a general motivational state that is revisable. An alternative is to suggest that there are different kinds of motivations—not just revisable preferences, but also inherited motivations like some emotions that cause agents to be more likely to undertake a kind of behavior. Thus, anger would be a disposition to retribute against the object of anger, fear a disposition to flee the feared object, and disgust a disposition to avoid or expel the disgusting object. Such an inherited motivation towards kinds of behavior can explain the individual mechanism for a general inherited strategy like retribution in the Reputation Game.

Some emotions, understood as motivations to kinds of behavior, can play an essential strategic role in some kinds of rational decision making *specifically because* they are motivations to a kind of behavior and not reducible to some combination of preferences or desires and other cognitive states. One way this can happen is to allow the emotions to overcome preferences that, on balance, should not take precedence, but which otherwise would. For example, disgust may need to motivate withdrawal and expelling of a toxin or pathogen directly, in order to overcome any preference that may arise from hunger. Fear may need to motivate flight directly, in order to overcome any preference that may be in competition with our assessment of the relevant danger. But there are also more complex strategic benefits that such kinds of motivations might serve in social interaction. In the Reputation Game, and even more so in a mix of Reputation and Ultimatum Games, a proven willingness to act retributively (that is, a high relative number of past rejections of proposals) benefits an individual agent by earning higher secondary offers.

As an example of such a strategic benefit, consider that anger, if it is a motivation to retribute, can serve to produce beneficially credible threats. Often, in social interactions, if an agent is unable to make a credible threat he will as a result earn a suboptimal outcome. Threats of the relevant kind often cannot be made if the decisions of the agent are based solely on expected utility calculated from minimal rationality and well ordered preferences. But if someone has a reputation for being willing to retribute even at personal cost, they can make a credible threat, and this can result in a strategic benefit in social interaction.

Human subject research on Ultimatum Games has consistently revealed behavior in which individuals reject low offers when acting as receivers, and make nearly even offers when acting as proposers. If human beings are motivated by anger to retribute when one gets a division that one perceives as unjust, then we have an explanation at the individual level for this outcome; the only way to retribute available to the receiver is to reject the proposal, and this is chosen often when proposals are low. In turn, if we all know this about human beings, we should be inclined to make proposals, when acting as proposers, which are less likely to anger the other player. Anger therefore also can explain why proposers typically make near-even division in their offers even in these one shot and anonymous games. Similar kinds of behavior can be observed in public goods games, where groups of individuals must decide upon common investments, and typically will resort to no investment, hurting everyone in the group, as a way to punish the free riders in the group (Andreoni 1995, Gintis et al. 2003). This, also, appears best explained in terms of a motivation to retribution.

A role for anger and other emotions in decision making, similar to what I am proposing here, was proposed by Robert Frank (1988). Frank argues that emotions cause certain feelings, which we prefer (or prefer not to have). These feelings in turn can serve to ensure certain kinds of commitments, such as credible threats. This reasoning is common in economics and game theory, and such preferences often are called "social preferences." Social preferences are preferences for a kind of outcome that is not directly or obviously of self-interest to the individual. However, there are three reasons, each independently sufficient, why understanding anger as a motivation to retribution, as opposed to reducing the role of anger to another preference, is a far superior explanation.

First, many of the behaviors motivated by anger and other kinds of emotions are not well explained by desires or preferences. There are many instances in which anger or fear motivate behaviors that extend beyond the satisfaction of any plausible preference the agent may have. These include kicking a tree when angry, or fleeing farther from a threat than one knows one needs to flee (see DeLancey 1998, 2002).

Second, if anger were just another preference, then it would allow for credible threats in fewer kinds of cases. A social preference invites calculation about the size of the social preference, and reduces the cases in which a credible threat can be made. Social preferences, if they are interpreted as real states (just like other preferences or desires) invite the other agent to take advantage in any situation where he believes he knows the magnitude of that social preference. We find ourselves quickly back in a situation like the original game, slightly reordered. But if irascibility is, per se, the motivation to retribute, then it is difficult for the other player in the game to calculate when threats become incredible, and thus it is difficult to know

when to take advantage. Irascibility bears an unpredictable relationship to other preferences, and for this reason being irascible makes almost any threat credible, even if unlikely.

Third, the inferential conditions of the threat of retribution in anger are straightforward. We have inherited from our biology and our culture a number of means of signaling whether we are irascible. We also have inherited from our biology and our culture a number of means of recognizing these signals. Furthermore, we all know (it is common knowledge) that all people are to some degree irascible, and we all learn that people seek to retribute when angry (so that even in anonymous interactions we recognize such a threat). This is a very minimal set of knowledge required for making a strategic evaluation. That is, in the social preferences explanation, for the threat to be credible one must know a number of things that it is difficult to know; one must know at least (1) that the other player understands the pay-offs in the game, (2) that he genuinely has the relevant social preference, (3) that the magnitude of this social preference exceeds his preferences to avoid the cost of exercising it, and (4) that he is rational. What the agent must know can in fact grow even more complex. If that agent begins to reason about the other player's understanding of his own reasoning, then he must consider whether the other agent knows what he knows, whether the other agent knows that he knows that that agent knows, and so on (see Bicchieri 1993). On the other hand, in the case of anger as a motivation to retribution, the task for the agent is arguably much simpler; he must determine whether it is likely that the other player is irascible.

Emotions as inherited, pancultural capabilities that provide revision-resistant motivations are a best explanation at the individual level of the mechanism involved in the kind of behavior observed in Ultimatum Games and related games. One of their principle roles is to allow for credible threats. There are many kinds of games where the ability to make a credible threat is essential to earning a better return in the game and where such better returns are observed and seem to be generated because of such threats. Anger, as a motivation to retribute as opposed to just another preference, is ideally suited to play this role.

It is plausible that other emotions also can play similar strategic roles. Fear as a motivation to flight, for example, might allow agents to make credible threats to defect in dangerous games of coordination, in turn forcing other players to bear greater costs. Consider, for example, the position of a Soldier given orders in battle, and the calculation of the commander who cannot assume (no matter what the preferences he ascribes to the Soldier) that all such orders will be followed. Fear allows for a credible threat of flight (avoidance of the task), even under the threat of other costs, such as long term or less dangerous (though not necessarily smaller) costs. This is a game where the options for the Soldier are obey or disobey, with respective outcomes. If threatened with disciplinary action for disobedience, the individual may fear the immediate danger far more than a distant threat of punishment (even if he prefers the danger to the punishment—the claim is not that fear is a preference, nor that it must be consistent with one's preferences), resulting in a significant revision-resistant motivation that could motivate seemingly irrational action. As is the case in the role of anger in enforcing credible threats of retaliation, such a credible threat of flight may have the tactical effect (and, from the perspective of some, benefit) of rebalancing the expected equilibrium in that particular game (e.g., options must be evaluated also in terms of their fearfulness).

Such examples establish that emotions as motivations to a kind of behavior can play a role in practical action of a kind that many forms of economic theory and game theory are concerned to explain, and that they can best play this role sometimes only because they are motivations to a kind of behavior as opposed to their motivational aspect being reducible to preferences or desires.

Conclusion

Life experience has made all of us generally quite good at predicting the kinds of situations that will make an individual fearful or angry. What we often appear to overlook is the important strategic role that such emotions can play. The concept of revision-resistance, and of approximate case games, can provide tools to conceive of such emotions in strategic interactions as a response that enforces or encourages certain forms of interaction. Such behaviors are far more likely than standard game theory or economics would predict, as confirmed by experimental economics. Revision-resistant motivations are inherited pancultural capabilities, and must be recognized not only for their seeming "irrational" influence but also for their strategic role and benefits. In the case of human performance in Ultimatum Games and related games, this provides a plausible explanation at the individual level of how strategies of reputation building and credible threats could be realized.

References

Andreoni, J. (1995). Cooperation in public goods experiments: Kindness or confusion. *American Economic Review,* 891-904.

Bicchieri, C. (1993). *Rationality and coordination (Cambridge studies in probability, induction, and decision theory).* New York: Cambridge University Press.

Bowles, S., & Gintis, H. (2002). The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical Population Biology, 65*, 17-28.

Camerer, C. (1995). Individual decision making. In J. Kagel & A. Roth (Eds.), *The handbook of experimental economics* (pp. 587-683). Princeton, NJ: Princeton University Press.

Camerer, C. (2003). *Behavioral game theory.* Princeton, NJ: Princeton University Press.

Camerer, C., & Thaler, R. (1995). Ultimatums, dictators, and manners. *Journal of Economic Perspectives 9*(2), 209-219.

DeLancey, C. (1998). Real emotions. *Philosophical Psychology,* 11:4.

DeLancey, C. (2002). *Passionate engines.* New York: Oxford University Press.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions.* New York: W. W. Norton & Company.

Gintis, H. (2000). *Game theory evolving.* Princeton, NJ: Princeton University Press.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24*, 153-172.

Guth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization, 3*, 367-388.

Hamilton, W. D. (1964). The genetical evolution of social behavior (I and II). *Journal of Theoretical Biology, 7*, 1-52.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (2004) *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies.* New York: Oxford University Press.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, G., McElreath, R. et al. (2005) "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28*: 795-855.

Hirshleifer, J. (August 1984). The emotions as guarantors of threats and promises. UCLA Department of Economics Working Paper.

Kovach, A., & DeLancey, C. (2005). On emotions and the explanation of behavior. *Nous,* 39(1): 106-122.

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science,* 207: 557-58.

Maynard Smith, J. (1982). *Evolution and the theory of games.* New York: Cambridge University Press.

Mitchell, M. (1998). *An introduction to genetic algorithms.* Boston: MIT Press.

Nowak, M. & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature 393*: 573-577.

Nowak, M. & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature 437*, 1291-1298.

Öhman, A. (1988). Preattentive processes in the generation of emotions. In V. Hamilton, G. H. Bower, & N. H. Frijda (Eds.), *Cognitive perspectives on emotion and motivation* (pp. 127-44). Boston: Kluwer Academic Publishers.

Öhman, A., Dimberg, U., & Esteves, F. (1989). Preattentive activation of aversive emotions. In T. Archer & L. G. Nilsson (Eds.), *Aversion, avoidance, and anxiety,* (pp. 169-93). Hillsdale, NJ: Lawrence Erlbaum.

Öhman, A. & Soares, J. J. F. (1993). On the automatic nature of phobic fear: conditioned electrodermal responses to masked fear-relevant stimuli. *Journal of Abnormal Psychology, 102*(1): 121-132.

Öhman, A. & Soares, J. J. F. (1994) "Unconscious anxiety": Phobic responses to masked stimuli. *Journal of Abnormal Psychology, 103*(2): 231-240.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica 50*: 97-110.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science, 300*, 1755-1758.

Selten, R. (1965). Spieltheoretische behandlung eines oligopolmodells mit nachfragetragheit. *Zeeitschrift fur die gesamte Staatswissenschaft, 121*, 301-324 and 667-689.

Sen, A. (2002). *Rationality and freedom.* Cambridge, MA: Harvard University Press.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monograph, 9*(2): 1-28.